纳米出版及其应用研究进展*

■ 牛丽慧 欧石燕

南京大学信息管理学院 南京 210023

摘要：［**目的/意义**］随着学术期刊文献的大量增长，在传统科学文献出版模式下，科研人员需要花费大量时间从文献中查找、获取和解读所需信息。为了促进科学信息的传播与交流，面向科学文献内容的细粒度语义出版成为一种新趋势。本文介绍语义出版中的一种代表性出版模式"纳米出版（nanopublication）"，并剖析纳米出版在不同学科领域中应用的可能性及应用特点。［**方法/过程**］首先对纳米出版模型进行了介绍，然后通过文献调研对纳米出版的应用现状进行了述评，最后以实例说明纳米出版在不同学科领域中的应用特点。［**结果/结论**］研究结果表明：①纳米出版目前主要应用于生物医学领域，在计算机和人文科学有少量应用，在其他领域几乎没有什么应用；②纳米出版可以扩展到其他学科领域进行应用，但是需要根据学科特征构建符合学科领域特点的纳米出版物。

关键词：纳米出版 语义出版 知识表示

分类号：G250

DOI：10.13266/j.issn.0252－3116.2018.07.015

1 引言

随着大数据时代的到来，用户可获取的信息资源越来越多，数字出版技术的发展更是导致了科学文献信息的爆炸式增长。在当前数字出版模式下，科学文献大多采用类似传统印刷型出版物的非结构化形式在线出版（如HTML、PDF格式），使得大量科学发现和科学结论被淹没在海量科学文本之中，科研人员需要花费大量时间查找文献、阅读并发现其中的关键信息，因此难以快速获取所需的科学信息以及其中的关联关系，为科学知识的交流、共享和重用带来了很大障碍。2009年牛津大学的D. Shotton教授首次正式提出了"语义出版（Semantic Publishing）"概念，受到学术界和出版界的广泛关注。D. Shotton认为语义出版是一种语义增强的期刊出版形式，采用语义网技术对期刊论文中的信息进行语义标注和语义关联，丰富出版物的内容，增强论文的语义，促进知识传播和学术交流[1]。基于语义出版思想，2009年非盈利性学术组织概念网络联盟（Concept Web Alliance）提出了一种新的科学信息出版模式——纳米出版（Nanopublication）[2]，旨在建立一种结构化和语义化的细粒度知识表示和发布方式。在纳米出版模式下，淹没在非结构化科学文本或海量数据集中的科学事实或科学结论被抽取出来，对其进行结构化和语义化表示，形成带有语境和支持信息的独立纳米出版物，一方面增强了科学信息的机器可读性，另一方面可促进科学知识的传播与交流。

本文首先对纳米出版物的结构、表示方式和实现步骤进行介绍和总结，然后对纳米出版的研究与发展现状进行分析与述评，最后通过实例分析比较了纳米出版在不同学科领域中应用的特点，为不同类型科学知识的细粒度语义化表示提供参考。

2 纳米出版物的结构与技术实现

为了解决在大数据背景下科学知识的发现、连接和监护问题，概念网络联盟于2009年5月首次提出了纳米出版物的概念，指出纳米出版物是可出版信息的最小单元，是能够被唯一标识且具有作者归属的关于任何事物的一个断言[3]。在此概念基础上，2010年荷

* 本文系国家社会科学基金重点项目"基于关联数据的学术文献内容语义发布及其应用研究"（项目编号：17ATQ001）和江苏省2016年度"青蓝工程"研究成果之一。

作者简介：牛丽慧（ORCID：0000－0003－3475－2860），博士研究生；欧石燕（ORCID：0000－0001－8617－6987），教授，博士生导师，通讯作者，E-mail：oushiyan@nju.edu.cn。

收稿日期：2017－09－16 修回日期：2017－11－30 本文起止页码：125－133 本文责任编辑：王善军

兰阿姆斯特丹自由大学的 P. Groth 等人提出了纳米出版模型,并初步提出了生成纳米出版物的技术规范[4]。2011 年 3 月欧洲创新药物计划资助的 Open PHACTS (Open Pharmacological Concept Triple Store)项目开始实施,采用纳米出版模式作为药物数据的统一表示形式以利于数据集成和互操作,从而为药物发现研究提供了语义基础设施,实现了纳米出版物的首次实际应用[5]。2012 年 6 月 Open PHACTS 项目发布了纳米出版物指南,定义了纳米出版物的组成元素和表示方法[6],在 2015 年发布了纳米出版物指南最新版本,同时创建了 nanopub. org 网站[3],有力地推动了纳米出版的发展。

2.1　纳米出版物的结构

根据概念网络联盟在 2015 年发布的最新版纳米出版物指南,纳米出版物的组成结构如图 1 所示:

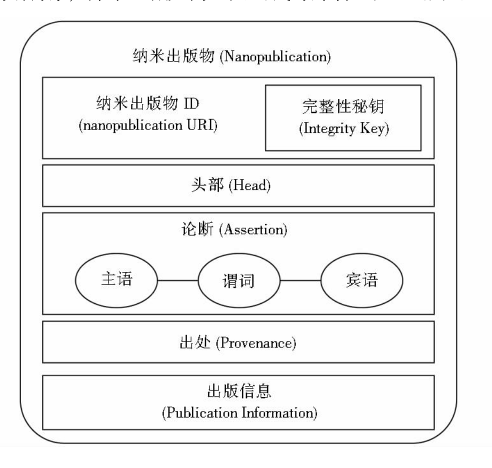


图 1　纳米出版物结构[3]

纳米出版物中包含三个基本组成部分[7]:①论断(Assertion):是最小的无歧义的信息单元,采用主－谓－宾三元组形式表示概念与概念之间的关系,是纳米出版物的核心,通常用于表示科学观点或科学结论;②出处(Provenance):表示论断的来源,包括提出论断的作者、机构、时间和地点等;③出版信息(Publication Information):即纳米出版物的元数据,包括该纳米出版物的创建者、创建时间、版权信息和版本信息等。

此外,纳米出版物还包括两个辅助要素:头部(Head)和纳米出版物 ID。其中,头部用于表明纳米出版物与论断、出处和出版信息之间的关系。纳米出版物 ID 是纳米出版物的唯一标识,用以保证纳米出版物的唯一性,通常采用 URI 标识符表示。完整性密钥(Integrity Key)是纳米出版物 ID 的一部分,用于核查纳米出版物是否发生变化,确保纳米出版物的稳定性和永久性。苏黎世联邦理工大学的学者 T. Kuhn 推荐使用可信任的 URIs(Trusty URIs)表示完整性密钥,即采用哈希加密算法对纳米出版物的内容添加哈希值以确保纳米出版物的可验证性和永久性[8]。

概念网络联盟采用 OWL 本体语言对上述纳米出版物结构进行了形式化描述,即构建了纳米出版物本体[8]。该本体主要包含一个纳米出版物类和三个 RDF 命名图(即论断图、出处图和出版信息图),以及描述它们之间的三个属性,如图 2 所示:

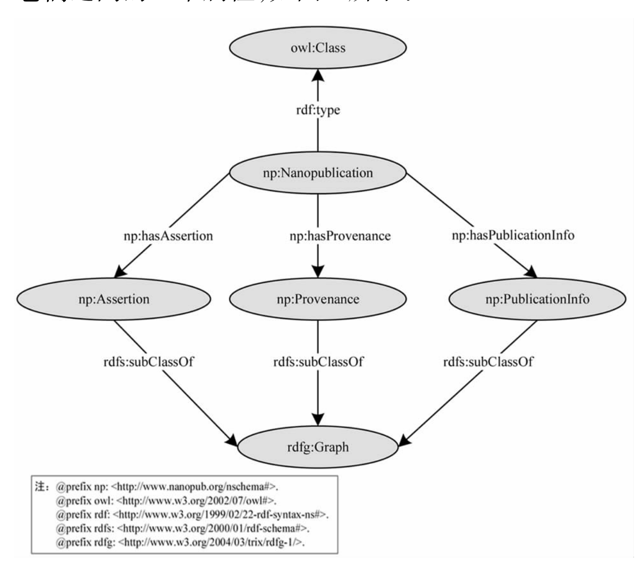


图 2　纳米出版物本体示意图

概念网络联盟采用 RDF 命名图(RDF Named Graph)对纳米出版物进行形式化表示[7]。RDF 命名图是对 RDF 三元组数据模型的一种简单扩展,是一组 RDF 三元组的集合,采用 URI 标识符进行标识,能够在其他 RDF 三元组中被直接引用作为主语或宾语[9]。命名图的序列化表示目前主要有三种:RDF/XML,TriX 和 TriG。其中 RDF/XML 和 TriX 是基于 XML 的序列化表示,而 TriG 则是 RDF/Turtle 格式的变形。纳米出版物的形式化表示由四个 RDF 命名图组成,即头部图(head graph),论断图(assertion graph),出处图(provenance graph)和出版信息图(publication information graph)。头部图包含四个 RDF 三元组,分别用于定义一个纳米出版物实例以及描述该实例与论断图、出处图和出版信息图之间的关系。论断图只能描述一个自然语言论断,需采用领域本体和专业词表对论断中的自然语言词汇进行规范化表示,可由一个或多个 RDF 三元组形式化地表示。出处图描述论断的背景信息(即论断的元数据),是对论断图属性的描述,由一个或多个主语是论断图的 RDF 三元组构成,谓语则常采用出处本体(Provenance Ontology)中的属性,譬如,

prov:genetatedAtTime 表示论断生成的时间,prov:wasDerivedFrom 表示论断的来源,prov:wasAttributiedTo 表示论断归属等[10];出版信息图描述纳米出版物的相关信息(即纳米出版物的元数据),由一个或多个主语是纳米出版物 URI 标识符的 RDF 三元组构成,描述属性可取自出处本体或相关元数据规范(如 DC 元数据)。

为了更加清晰地说明纳米出版物的结构,我们以生物医学领域中的一个科学论断“Trastuzumab is indicated for breast cancer”(曲妥单抗用于治疗乳腺癌)为例,将其表示为纳米出版物,如图 3 所示。在该图中,论断描述了“ex:trastuzumab”和“ex:breast-cancer”两个实体及其之间“ex:is-indicated-for”的关系;出处信息描述了该论断生成于 2012 年 2 月 3 日 14 点 38 分,来源于某个实验结论,归属于某个实验科学家;出版信息描述了该纳米出版物生成于 2012 年 10 月 26 日 12 点 45 分,创建者是 P. Groth。

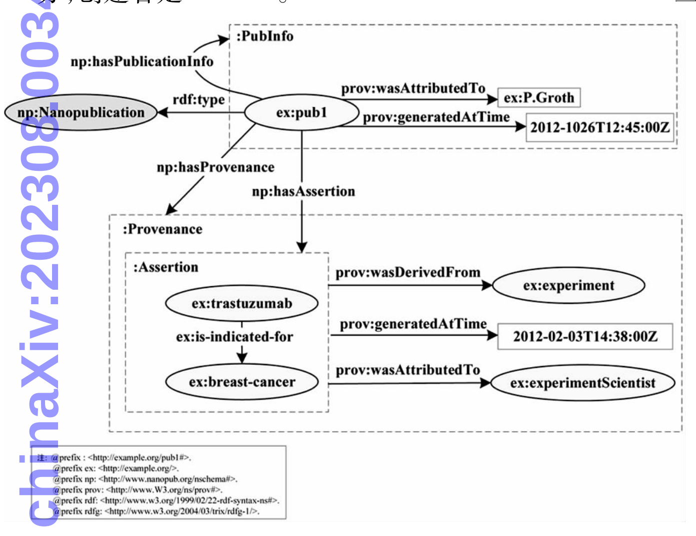


图 3　纳米出版物实例示意图

2.2　纳米出版物的技术实现

纳米出版物的技术实现主要包括以下四个步骤:

(1)科学论断的抽取:构建纳米出版的首要工作是从非结构化的科学文献中识别并抽取出重要的科学观点或结论,将其作为纳米出版物的论断。鉴于科学文献通常具有比较固定的篇章结构,且科学观点或科学结论通常出现在科学文献的特定位置(如结论),因此通过对科学文献进行篇章结构解析,有助于快速定位重要的科学论断。关于科学文献的篇章结构目前有多个模型,其中一个代表性的模型是英国阿伯里斯特威斯大学开发的用于描述有实验方法驱动的科学研究的元数据模型 CISP(Core Information about Scientific Papers)[11],该模型将科学研究的主要方面划分为八个类别,分别为:研究目的、动机、研究对象、研究方法、实验过程、观察、实验结果和结论。CISP 的制定者还开发了基于 CISP 的辅助人工标注工具 SAPIENT(Semantic Annotation of Papers: Interface & Enrichment Tool),可用于辅助人工识别科学文献中表示上述八个关键类的自然句[12]。

(2)科学论断的语义化表示:将从文献中抽取出的自然语言论断转换为一组 RDF 三元组表示。首先,通过自然语言处理技术将抽取出的自然语句转换成一个主－谓－宾三元组形式的逻辑表示,逻辑三元组中的主语和宾语分别对应于自然句中作为主语和宾语的名词或名词短语,谓语则对应于自然句中描述两者之间关系的词汇(如动词短语、介词短语)。其次,需要将逻辑三元组中的自然语言词汇转换为规范词汇,并采用 URI 标识符进行唯一标识,才能将其转换为 RDF 三元组。鉴于纳米出版模型中并没有提供用于规范化表示的词表或本体,因此可重用已有的开放公共词表或本体,而且这更有助于纳米出版物与其他资源信息的集成。Open PHACTS 组织推荐了一些常用词表用于描述生命科学领域的概念,如表 1 所示[6]。为了对 RDF 三元组中规范化表示的实体或概念进行唯一标识,一些学术机构专门建立了提供 URIs 标识的服务系统,如欧洲生物信息研究所为共享生命科学实验数据而构建的一个维护数据唯一性的服务系统 Identifiers. org,该系统主要为生命科学领域的实体对象提供 URIs,其形式为 <http://identifiers. org/[namespace]/[entity]>[13]。最后将转化成的 RDF 三元组以 RDF 命名图的形式表示,作为纳米出版物的论断图。

(3)科学论断出处信息的构建:为来自科学文献的论断添加元数据,主要包含论断的作者、机构、时间和地点信息等。在构建纳米出版物时,主要采用出处本体中的属性来描述上述元数据。形式化表示的论断元数据(即出处信息)构成了一个 RDF 命名图,称作纳米出版物的出处图。

(4)纳米出版物出版信息的构建:为形式化表示的纳米出版物添加元数据,主要包含纳米出版物的创建者、创建时间、版权和版本信息等,为纳米出版物提供背景和语境信息。在构建纳米出版物时,主要采用 DC 核心元数据和出处本体中的属性来描述。形式化表示的纳米出版物元数据(即出版信息)构成了一个 RDF 命名图,称作纳米出版物的出版信息图。

表 1　Open PHACTS 组织推荐的常用词表

词表名称	词表覆盖领域	词表 URI 标识
DC 核心元数据	通用	http://dublincore. org/
SWAN 本体	轻量级出处	http://swan. mindinformatics. org/
开放出处模型	出处信息	http://openprovenance. org/
纳米出版物本体	纳米出版物概念	http://www. nanopub. org/nschema
出版角色本体	出版概念	http://vocab. ox. ac. uk/pro
数据立方体	统计学	http://publishing-statistical-data. googlecode. com/svn/trunk/specs/src/main/html/cube. html
实验因素本体	通用的实验概念	http://www. ebi. ac. uk/efo/
创作共享词表	许可信息	http://wiki. creativecommons. org/CC_REL
语义科学本体	通用语义关系的谓词	http://semanticscience. org/ontology/sio-core. owl

3　纳米出版的应用现状

目前,我们采用 Google Scholar 搜索引擎从网络上共收集到有关纳米出版的相关论文 66 篇,其中 22 篇是关于纳米出版概念或模型本身的阐述,其余 44 篇是关于纳米出版的实际应用。我们对这 44 篇文献的应用领域进行了分析,发现纳米出版主要应用于生物医学、计算机和人文科学三个领域,具体分布如图 4 所示:

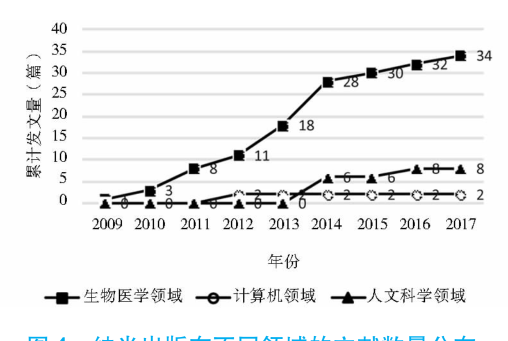


图 4　纳米出版在不同领域的文献数量分布

从图 4 可以看出,纳米出版在生物医学领域的应用相对比较成熟,占据了目前 80% 左右的应用。从发布为纳米出版物的数据来源来看,主要分为两大类:一类是科研工作者自行将个人研究成果发布为纳米出版物,旨在促进最新研究成果的传播与交流;另一类是将已有关系型数据库中的数据以纳米出版形式发布,旨在发现数据之间新的关联关系以及提高数据的可信度。在第一类应用中,一个具有代表性的案例是 2015 年荷兰莱顿大学的 E. Mina 等人采用纳米出版模型将关于亨廷顿病(一种显性遗传性神经系统变性病)的实验结果和实验过程发布为关联数据,从而弥补了传统期刊出版方式无法将实验数据展示给读者的缺点,并且增强了数据之间的关联性和互操作性。E. Mina 首先将科学文献中以自然语言表达的实验结果转化为 RDF 三元组,以此作为纳米出版物的论断,然后基于以工作流为中心的研究对象模型(Workflow-centric Research Object Model,简称 WROM)标注实验过程和实验数据,以此作为论断的出处信息[14]。针对第二类应用,一个代表性应用是 2005 年实施的欧盟创新医学联合计划项目 Open PHACTS。该项目将纳米出版应用到面向药物发现研究的语义基础设施中,采用纳米出版模式表示生物活性数据及结论性信息,构建一个药物数据集成平台,一方面提高数据的可信度和引用率,另一方面促进数据集成和互操作[15]。另一个代表性应用是 2017 年美国伦斯勒理工学院的 J. P. Mccusker 等人采用纳米出版模式对来自不同数据库中的药物－靶点、蛋白质－蛋白质、疾病－基因间关联关系进行统一的语义化表示和集中存储,通过这种方式解决了数据的碎片化问题,为药物再利用提供了数据支撑[16]。第二类应用的实现需要依赖于各种 RDF 自动转换工具,将不同领域和格式的结构化数据自动转换为 RDF 格式表示的纳米出版物。譬如,美国伦斯勒理工学院的 J. P. Mccusker 等人于 2013 年开发了 Prizms 工具,能够将 CSV、XML、JSON 等格式的数据自动转化为 RDF 格式[17];希腊佩特雷大学的 G. P. Patrinos 于 2012 年面向 LOVD 数据库(Leiden Open-Access Variation Database,莱顿开放获取基因变异数据库)开发了纳米出版工具,用于将人类基因变异数据发布为纳米出版物,以达到科学共享的目的[18]。

除了生物医学领域外,一些研究者也对纳米出版在其他领域的应用进行了探索。2014 年,奥地利维也纳理工大学的 A. Lipani 等人基于信息检索本体对描述信息检索实验的科学文献进行了语义标注,从而将信息检索实验数据以纳米出版物形式发布,以增强检索实验的可重复性[19]。2015 年,北卡罗来纳大学教堂山分校的 P. Golden 等人采用纳米出版模式对考古、历史和艺术史中有关历史时期的描述进行形式化表示,

并发布为关联数据[20]。2015年，马里兰大学人文科技学院的R. Viglianti用纳米出版模式将传统乐谱发布为关联数据，增强了音乐记谱法的寻址能力[21]。

自2013年起，国内也有学者开始对纳米出版进行研究和探讨，但总体仍处于起步阶段。中科院文献情报中心的吴思竹等于2013年最早在国内对纳米出版这一语义出版模式进行了介绍，并对纳米出版的概念、源文件、语义模式、结构描述、应用价值、在知识发现中的作用等方面进行了理论探讨[22]。随后，山西医科大学图书馆的苏云梅和武建光于2015年探讨了纳米出版与知识发现之间的关系，并以人类遗传疾病塞克尔综合征为例分析了纳米出版语义关联模式的偏差、不足及其原因[23-24]，但没有进行实践应用。上海师范大学人文与传播学院的吕元智于2015年构建了基于纳米出版模式的数字档案资源的语义描述框架[25]，旨在对数字档案资源的外部结构进行粗粒度描述，如档案提名、分类号、主题词等，这种应用方式与纳米出版面向细粒度知识表示的宗旨有很大偏差，不能看作是一个真正的纳米出版应用。

从纳米出版的发展及应用现状来看，纳米出版最早诞生于生物医学领域，并在该领域有较成熟的应用。这是因为：一方面，生物医学领域往往需要连接、整合数以万计的实验结果，用于验证实验结果或指导新的科学发现，而将研究结果以纳米出版物形式发布能够满足上述需求并有利于数据之间关联关系的发现；另一方面，生物医学领域的科学文献通常结构比较严谨，对实验过程及其结论的描述非常严格规范，且该领域对于规范化词汇的要求较高，拥有大量专门词汇和领域本体，因此将科学文献中的自然语言论断转换为结构化表示相对比较容易。纳米出版在非生物医学领域的应用拓展了纳米出版的应用范围，目前主要应用在计算机科学、历史和艺术学等学科领域，但在其他学科的应用仍是空白。若要达到不同学科间的知识共享和交流目的，需要将所有学科的知识都进行结构化和形式化表示，因此将纳米出版的应用拓展到其他学科领域有助于推动跨学科科学交流和知识共享。从目前国内对纳米出版的研究情况来看，主要集中在对纳米出版应用价值的探讨和纳米出版在一些领域中应用框架的构建，但缺乏对纳米出版的实际应用的研究实现。因此，如何将纳米出版模式应用于不同学科领域的科学文献或数据库中，对科学知识细粒度表示还有很大的探索空间。

4 纳米出版在不同学科领域应用特点剖析

随着各学科领域各种科学结论和数据的急速增加，以及对其进行自动语义集成、关联和推理的迫切需求，对纳米出版的应用进行拓展是十分必要的。苏黎世联邦理工大学的T. Kuhn等人认为，将所有科学数据都以纳米出版形式表示似乎是不太可能的，甚至会限制实际的应用范围[26]。基于此，北卡罗来纳大学教堂山分校的P. Golden等人提出，纳米出版物的论断应该根据使用它们的研究者的实际需要来进行构建，只要实际上能对目标用户有用即可，无需将数据严格统一化，否则会造成数据过度复杂，反而给用户的使用带来困难[20]。

纳米出版物的核心在于论断，而论断通常源自于科学文献中的声明。概念网络联盟成员B. Mons将科学文献中的声明分为三类：事实性声明(curated statements)、观察性声明(observational statements)和假设性声明(hypothetical statements)。其中，事实性声明是指经过专业审核校对过的科学事实，一般存储于专业数据库中，如在线人类孟德尔遗传数据库(OMIM)和蛋白质数据库(UniProt)，是构建领域本体的基础。观察性声明是指经过实验或统计得出的结论性声明，这类声明没有固定的数据库进行存储和管理。假设性声明是根据现有知识通过文本挖掘或直接推理而得，将这类声明进行聚合是知识发现的源泉[27]。不同学科领域对知识的描述方式和利用程度有很大不同，抽取的论断内容也有所不同。下面分别以生物医学、历史和社会学领域的科学文献为例，展示不同学科领域科学论断不同的特点，说明如何采用纳米出版模型将不同类型的科学论断表示为纳米出版物。

4.1 生物医学领域科学论断的特点与形式化表示

生物医学文献中的科学声明多为通过科学实验得出的科学结论，多表现为相关关系，通常位于文献的结论部分，并在文献的摘要部分也有所提及。譬如，在华盛顿大学P. Liu等人于2010年发表的一篇科学论文中，作者采用横断面研究方法探索了亚甲基四氢叶酸还原酶基因多态性与白人女性月经初潮年龄和自然绝经年龄的关系。该论文的摘要部分包含如下论断：

"The results of our study suggest that the MTHFR gene may influence the onset of menarche and natural menopause"(亚甲基四氢叶酸还原酶基因影响月经初潮和自然绝经)[28]。

该论断首先被简化为一个逻辑表示，即“MTHFR→influences→ the onset ofmenarche and natural menopause”。然后，采用国家癌症研究所词表（National Cancer Institute thesaurus，简称 NCIT）中的规范词汇“ncit：MethylenetetrahydrofolateReductase”描述基因“MTHFR”（亚甲基四氢叶酸还原酶），采用管理活动医学词典（Medical Dictionary for Regulatory Activities，简称 MEDDRA）中的规范词汇“meddra：Menarche”和“meddra：NaturalMenopause”描述“Menarche”（经初期）和“Natural Menopause”（自然绝经）这两种生理现象，采用 Dbpedia 本体中的属性“dbpedia：influence”表示“MTHFR gene”对“Menarche”和“Natural Menopause”的影响关系。最后，采用规范词汇代替逻辑表示中的自然语言词汇，将其转换为两个 RDF 三元组表示，作为纳米出版物中的论断，如图 5 所示：

```
@prefix : <http://www.example.org/example>.
@prefix dbpedia: <http://dbpedia.org/resource/>.
@prefix meddra: <http://purl.bioontology.org/ontology/MEDDRA/>.
@prefix ncit: <http://purl.bioontology.org/ontology/NCIT>.
:Assertion{
ncit:MethyleneterahydrofolateReductase
    dbpedia:influenced    meddra:Menarche,
                          meddra:NaturalMenopause.}
```

图 5　“基因影响经期”纳米出版物论断（Turtle 格式）

4.2　历史学领域科学论断的特点与形式化表示

与自然科学领域不同，人文科学领域的许多科学声明是作者的经验性观点或假设，通常是无法用实验验证和证伪的。因此，对于人文领域的科学文献，主要抽取以理论论述方式得出的观点或结论性知识，这类知识通常散落于科学文献的不同章节，多表现为对某一事物特点或属性的描述。譬如，在法国国家科学研究中心 P. Beaujard 于 2010 年发表的一篇科学论文中，作者提出一个观点，其文本内容为：

“In China, the political fragmentation characterizing the period known as Spring-and-Autumn (771BC-481BC) went along with the development of the craft industry and exchange”（在中国，以政治分裂为特征的春秋时期（公元前 711 年 – 公元前 481 年）伴随手工业的发展和交流）[29]。

从该段文本中可提取出关于春秋时期时间跨度的描述作为纳米出版物的论断内容，即“China”（中国）的一个历史时期“Spring-and-Autumn”（春秋时期）起止于“771BC”和“481BC”。在该论断中，将时间实体“Spring-and-Autumn”表示为 DBpedia 本体中“dbpedia：SpringAndAutumnPeriod”类的实例，将国家“China”表示为“dbpedia：China”类的实例，采用 DC 核心元数据的属性“dcterms：spatial”表示春秋时期与中国的时空关系，采用时间本体中的属性“time：intervalStartedBy”和“time：intervalFinishedBy”表示“Spring-and-Autumn”的起止时间，采用 SKOS 模型中的属性“skos：prefLabel”表示以公元纪年“771BC”和“481BC”表示的春秋时期起止时间。由此，将该论断转化为 5 个 RDF 三元组，以此作为纳米出版物的论断，如图 6 所示：

```
@prefix  :  <http://www.example.org/example>.
@prefix  dbpedia:  <http://dbpedia.org/resource/>.
@prefix  skos:  <http://www.w3.org/2004/02/skos/core#>.
@prefix  time:  <http://www.w3.org/2006/time#>.
@prefix  dcterms:  <http://purl.org/dc/terms/>.
:Assertion{
:Spring-and-Autume    a    dbpedia:SpringAndAutumnPeriod;
      dcterms:spatial    :China;
      time:intervalStartedBy  [skos:prefLabel    "771BC"].
      time:intervalFinishedBy  [skos:prefLabel    "481BC"].
:China    a    dbpedia:China.}
```

图 6　“春秋历史时期时间跨度”纳米出版物论断（Turtle 格式）

4.3　社会学领域中科学论断的特点与形式化表示

社会科学领域中的科学文献主要有两大类：以逻辑论证为基础的理论型文献和以数据为基础的实验型文献。理论型文献的论述结构较为多样，没有统一的、较为固定的篇章结构，其论证结论一般为述评或建议等，其逻辑关系较为复杂，难以简化为三元组形式。此外，纳米出版的目的是将细粒度科学知识以结构化形式出版，构成一个网状的知识图谱，这些知识通常都是比较客观的科学事实或结论，而述评或建议的主观性往往较强，转化为纳米出版物的意义不大。相反，以数据为基础的实验型文献论述结构固定，实验结论较为客观，并且其结论相对较容易转化为三元组形式。因此，在社会科学领域，主要是将实验型文献中的科学结论发布为纳米出版物，这里科学结论通常位于文献的结论部分，并在文献的摘要部分也有所提及。譬如，在美国明德学院的 M. Lawrence 于 2016 年发表的一篇学术论文中，作者采用因果推断实验方法探索了父母受过高等教育情况与子女受高等教育情况两者之间的关系。该论文的摘要包含如下论断：

“Results show that college increases male graduates' probability of having a child who completes college”（结果表明，大学男性本科毕业生的子女完成大学学业的几率更高）[30]。

该论断可被简化成两个逻辑表示：①College→in-

creases→ probability;②The probability→refers → male college graduates' one child completes college。

在该论断中，我们采用心理学本体（Psychology Ontology，简称 APAONTO）、DBpedia 本体、知识库序列本体（SequenceOntology，简称 SO）和医学主题词表（Medical Subject Headings，简称 MESH）对该结论中的实体和关系进行规范化表示。由于在心理学本体中没有找到直接对应"male college graduates"的概念，我们对该本体进行扩展，为已有类"apanoto:CollegeGraduates"新增两个新的子类"MaleCollegeGraduates"和"FemaleCollegeGraduates"。然后，将实体"college"表示为心理学本体中类"apaonto:College"的实例，将实体"male college graduates"表示为新增类"apaonto:MaleCollegeGradutes"的实例。采用医学主题词表中的规范词汇"mesh:Probability"和"mesh:Child"分别表示名词"probability"和"child"。最后，将上述两个逻辑表示转换为一组 RDF 三元组，作为纳米出版物的论断。图7 为对上述结论使用 Turtle 格式表示的纳米出版物论断。

```
@prefix : <http://www.example.org/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix mesh: <http://purl.bioontology.org/ontology/MESH>.
@prefix dbpedia: <http://dbpedia.org/resource/>.
@prefix apaonto: <http://purl.bioontology.org/ontology/APAONTO>.
@prefix so: <http://www.sequenceontology.org/>.
:Assertion{
    :College  a  apaonto:College;
    dbpedia:increase  mesh:Probability.
    mesh:Probability  so:refers_to  :blank_node.
    :blank_node  a  rdf:Statement.
    :blank_node  rdf:subject  :MaleCollegeGraduates.
    :blank_node  rdf:predicate  dbpedia:has_a.
    :blank_node  rdf:object  mesh:Child.
    mesh:Child  dbpedia:complete  :College.
:MaleCollegeGraduates  a  apaonto:MaleCollegeGraduates.}
```

图7　社会科学领域纳米出版物论断（Turtle 格式）

通过对纳米出版物在上述三个不同学科领域的应用实践，我们认为纳米出版在不同学科领域的应用具有以下几个特点：

（1）不同学科领域中可表示为纳米出版物的科学论断类型不同。以实验为基础的学科领域中可表示为纳米出版物的有事实性声明和观察性声明，该类声明一般位于科学文献的摘要或结论部分，可直接锁定这两部分进行论断的抽取。非实验性学科领域中可表示为纳米出版物的通常是作者的观点或公认的共识，这类声明在科学文献中的位置不确定，需要根据实际情况设置线索词进行论断的抽取。

（2）不同学科领域中将科学结论表示为纳米出版物的精确度不同。自然科学与人文科学最大的区别在于：前者是对自然客观事实的认识，后者是对社会现象的认识。在自然科学领域，大部分实验结论是不受主观影响的，对事实的反映是唯一确定并可重复性验证的，因此自然科学领域中的纳米出版物通常具有一致性，很少存在冲突矛盾。与自然科学领域不同，人文科学领域产生的许多科学声明受主体文化背景影响较大，对同一现象的理解存在一定偏差，因此对人文科学领域中的观点和结论采用纳米出版物表示时需要有一定的容差，对结论限制得太精确反而会影响实际应用效果。

（3）不同学科领域中将科学论断发布为纳米出版物的用途不同。在自然科学领域，科学知识更新速度较快，产生的科学成果数量也较多，给科研工作者在短时间内查阅资料带来困扰。纳米出版物是最小出版单元，是适应于大数据时代研究成果快速出版的一种出版方式，有利于自然科学领域研究成果的快速发布。科学结论的纳米出版物化有利于验证实验结果和方法的有效性以及促进知识共享和重用。在人文和社会科学领域，纳米出版可作为文化遗产数据、金融数据、医疗与健康数据等多种类型数据的统一出版方式，以实现不同领域的知识关联，进而促进隐性知识显性化。此外，人文社科领域研究范式的转变需要数据质量保障机制，纳米出版物的每一个论断都附带有出处信息，将数据表示为纳米出版物保障了数据的质量和可信度。

鉴于此，在对不同学科的科学声明或结论使用纳米出版进行语义化表示时，需要善于捕捉各学科的特点，构建符合该学科领域特征的纳米出版物。

5　结论与展望

纳米出版的提出实现了科研信息从全文期刊形式发布到细粒度知识单元形式发布的转换，将科学事实和科学结论以结构化形式独立发布，用出处信息表示结论的来源增加其信任度，用 URI 对结论进行唯一标识，任何人都可通过网络对其进行检索和引用，使得研究成果的被引用率不再受限于期刊的知名度，因此在一定程度上可缩小不同地区的研究者因基础设施或地区的差别而对科学做出贡献的差距。

虽然纳米出版的意义不可否认，但同时也存在一些缺陷：①将科学结论以纳米出版形式发布需要作者自行构建，但通常情况下作者习惯于采用自然语言形

式表达他们的研究结论,而且相对于形式化结论带来的益处,构建纳米出版物过程的繁琐对于普通科学工作者而言问题更加严重;②纳米出版无法解决论断的冲突问题,纳米出版物只能如实将科学文献的结论进行语义表达,但当两个作者提出的观点存在冲突时,无法通过纳米出版物辨别哪个结论是正确的;③纳米出版主要是把科学文献中的结论性声明进行形式化表示,但对于文献中通过引文、实验数据或实验方法论证作者观点的过程无法形式化描述,因此无法根据纳米出版物对实验结果进行二次重复。

在下一步工作中,我们拟探索纳米出版在社会科学领域中的应用,以期拓展纳米出版的应用范围。鉴于纳米出版本身存在的缺陷,我们拟结合其他语义出版模型构造一个新的语义出版模型,不仅能对科学文献中的科学结论进行语义化表示,还能对科学结论的原文出处、论证过程等进行语义化表示,以提高纳米出版物的可信度和有效性。

参考文献:

[1] SHOTTON D. Semantic publishing: the coming revolution in scientific journal publishing[J]. Learned publishing, 2009, 22(2): 85–94.
[2] Concept Web Alliance[EB/OL]. [2017–07–16]. https://www.nbic.nl/about-nbic/affiliated-organisations/cwa/introduction/index.html.
[3] Nanopub.org[EB/OL]. [2017–07–16]. http://nanopub.org/wordpress/.
[4] GROTH P, GIBSON A, VELTEROP J. The anatomy of a nanopublication[J]. Information services & use, 2010, 30(1/2): 51–56.
[5] Open PHACTS[EB/OL]. [2017–07–16]. http://www.openphacts.org/index.php.
[6] The open PHACTS nanopublication guidelines[EB/OL]. [2017–07–16]. http://www.nanopub.org/guidelines/1.8/.
[7] Nanopublication guidelines[EB/OL]. [2017–07–16]. http://nanopub.org/guidelines/working_draft/.
[8] KUHNT. Nanopub-java: a java library for nanopublications[EB/OL]. [2017–07–16]. http://dare.ubvu.vu.nl/handle/1871/53984.
[9] BIZER C, CARROLL J J, HAYES P, et al. Named graphs, provenance and trust[EB/OL]. [2017–07–16]. http://lists.w3.org/Archives/Public/www-archive/2004Mar/att-0118/ng.pdf.
[10] PROV-O: the PROV ontology[EB/OL]. [2017–07–16]. http://www.w3.org/TR/prov-o/.
[11] SOLDATOVA L, LIAKATA M. An ontology methodology and CISP-the proposed core information about scientific papers[EB/OL]. [2017–07–16]. http://www.aber.ac.uk/en/media/departmental/impacs/computerscience/pdfs/ReportCISPshort.pdf.
[12] LIAKATA M, CLAIRE Q, SOLDATOVA L N. Semantic annotation of papers: interface & enrichment tool (sapient)[EB/OL]. [2017–07–16]. http://www.aber.ac.uk/en/media/departmental/impacs/computerscience/pdfs/sapientBio2009Final.pdf.
[13] Identifiers.org[EB/OL]. [2017–07–16]. http://identifiers.org/.
[14] MINA E, THOMPSON M, KALIYAPERUMAL R, et al. Nanopublications for exposing experimental data in the life-sciences: a huntington's disease case study[J]. Journal of biomedical semantics, 2015, 6(1): 1–12.
[15] HARLAND L. Open PHACTS: a semantic knowledge infrastructure for public and commercial drug discovery research[EB/OL]. [2017–07–16]. http://www.openphacts.org/documents/publitions/121008_EKAW_Lee%20Harland_A%20semantic%20infrastructure%20for%20public%20and%20commercial%20drug%20discovery.pdf.
[16] MCCUSKER J P, DUMONTIER M, YAN R, et al. Finding melanoma drugs through a probabilistic knowledge graph[EB/OL]. [2017–07–16]. https://peerj.com/articles/cs-106.pdf.
[17] MCCUSKER J P, LEBO T, KRAUTHAMMER M, et al. Next generation cancer data discovery, access, and integration using prizms and nanopublications[EB/OL]. [2017–07–16]. http://pdfs.semanticscholar.org/92b0/b95eedac4cd3cbbb1e7ff73900f94ceda21e.pdf.
[18] PATRINOS G P, COOPER D N, MULLIGEN E V, et al. Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain[J]. Human mutation, 2012, 33(11): 1–10.
[19] LIPANI A, PIROI F, ANDERSSON L, et al. Extracting nanopublications from IR papers[EB/OL]. [2017–07–16]. https://www.researchgate.net/publication/267402028_Extracting_Nanopublications_from_IR_Papers.
[20] GOLDEN P, SHAW R. Nanopublication beyond the sciences[EB/OL]. [2017–07–16]. https://dx.doi.org/10.7287/peerj.preprints.1284v1.
[21] Enhancing music notation addressability[EB/OL]. [2017–07–16]. http://mith.umd.edu/research/enhancing-music-notation-addressability/.
[22] 吴思竹,李峰,张智雄. 知识资源的语义表示和出版模式研究——以 Nanopublication 为例[J]. 中国图书馆学报,2013(4):102–109.
[23] 苏云梅,武建光. 纳米出版物语义模式及其与知识发现的关系[J]. 中华医学图书情报杂志,2015(12):15–18.
[24] 武建光,苏云梅. 基于知识发现的 Nanopublication 语义模式研究[J]. 山西医科大学学报,2015(8):833–836.
[25] 吕元智. 基于 Nanopublication 框架的数字档案资源语义描述研

究[J]. 档案学通讯,2015(3):57 – 62.
[26] KUHN T, BARBANO P E, NAGY M L, et al. Broadening the scope of nanopublications[J]. Lecture notes in computer science, 2013,7882:487 – 501.
[27] MONS B, VELTEROP J. Nano-Publication in the e-science era [EB/OL]. [2017 – 07 – 16]. https://www.w3.org/wiki/images/4/4a/HCLS$$ISWC2009$$Workshop$Mons.pdf.
[28] LIU P, LU Y, RECKER R R, et al. Association analyses suggest multiple interaction effects of the methylenetetrahydrofolate reductase polymorphisms on timing of menarche and natural menopause in white women[J]. Menopause,2010,17(1): 185 – 190.
[29] PHILIPPE B. From three possible Iron-age world-systems to a single Afro-Eurasian world-system[J]. Journal of world history, 2010,21(1): 1 – 43.
[30] LAWRENCE M, RICHARD B. And their children after them? The effect of college on educational reproduction[J]. American journal of sociology, 2016,122(2): 532 – 572.

作者贡献说明:
牛丽慧:负责资料收集,论文撰写与修改;
欧石燕:指出研究方向和研究思路,论文修改、审阅和定稿。

Research Advances of Nanopublication and Its Applications

Niu Lihui　Ou Shiyan
School of Information Management, Nanjing University, Nanjing 210023

Abstract: [**Purpose/significance**] With the substantial growth of the number of academic articles, researchers usually spend more time in searching, acquiring and understanding the content of academic articles under the current publishing modes. In order to facilitate scientific information can be disseminated and exchanged rapidly, a new publishing mode: semantic publishing, which focuses more on the fine-grained content of academic journals, is arising in academic communities. This paper intends to introduce a representative semantic publishing mode: "nanopublication", and explore the possibility and characteristics of application in different subject fields. [**Method/process**] Firstly, we introduced the nanopublication model. Secondly, we reviewed the current status of nanopublication's application by literature review. Finally, we analyzed the characteristics of nanopublication's application in different subject areas with examples. [**Result/conclusion**] The research shows that: ①Nanopublication is mainly used in biomedical field currently, seldom used in the fields of computer science and human science, and hardly used in other fields; ②It is possible to extend nanopublication to other subject fields, and construct nanopublications according to the characteristics of different subject fields.

Keywords: nanopublication　semantic publishing　knowledge representation